DOCUMENT RESUME

ED 093 994 TM 003 850

AUTHOR Aleamoni, Lawrence M.
TITLE The Relation of Sample Size to the Number of Variables in Using Factor Analysis Techniques.
INSTITUTION Illinois Univ., Urbana. Office of Instructional Resources.
REPORT NO TR-4
PUB DATE Feb 74
NOTE 8p.

EDRS PRICE MF-$0.75 HC-$1.50 PLUS POSTAGE
DESCRIPTORS Correlation; *Factor Analysis; *Factor Structure; *Matrices; *Sampling; Statistical Bias

ABSTRACT

The relationship of sample size to number of variables in the use of factor analysis has been treated by many investigators. In attempting to explore what the minimum sample size should be, none of these investigators pointed out the constraints imposed on the dimensionality of the variables by using a sample size smaller than the number of variables. A review of studies in this area is made as well as suggestions for resolution of the problem. (Author)

Technical Report

No. 4

TITLE The Relation of Sample Size to the Number of Variables in Using Factor Analysis Techniques

AUTHOR(S) Lawrence M. Aleamoni

DATE February, 1974

MEASUREMENT AND RESEARCH DIVISION - OFFICE OF INSTRUCTIONAL RESOURCES
307 ENGINEERING HALL - UNIVERSITY OF ILLINOIS
URBANA, ILLINOIS 61801

Abstract

The relationship of sample size to number of variables in the use of factor analysis has been treated by many investigators. In attempting to explore what the minimum sample size should be, none of these investigators pointed out the constraints imposed on the dimensionality of the variables by using a sample size smaller than the number of variables. A review of studies in this area is made as well as suggestions for resolution of the problem.

THE RELATION OF SAMPLE SIZE TO THE NUMBER OF VARIABLES IN USING FACTOR ANALYSIS TECHNIQUES

Lawrence M. Aleamoni[1]

One of the basic axioms of factor analysis is that the number of variables (V) in the correlation matrix should not exceed the number of observations (N). In fact, this axiom is so taken for granted by Thurstone (1947), Guilford (1954), and Harman (1960) that they state it without any supporting reasons.

Of course, the most obvious reason for having N greater than V is that otherwise one restricts the maximum number of linearly independent factors that can be extracted from the correlation matrix. This can readily be shown by the following:

> *The rank of a matrix is the maximum number of linearly independent row (or column) vectors of that matrix (Murdoch, 1957). The rank of the product of two matrices is less than or equal to the rank of either matrix. Whenever the number of rows is not equal to the number of columns of a matrix the maximum number of linearly independent vectors is equal to the smaller number of rows or columns of that matrix. Thus, the maximum number of linearly independent factors that can be extracted from a correlation matrix R is equal to the rank of either matrix that is used to generate R.*

Now the correlation matrix R equals FF' and by a well-known theorem, the rank of R is less than or equal to the rank of F' or F, whichever is smaller. But F and F' are mutual transposes and so their ranks are equal and, therefore, the rank of F equals the rank of R (for example, see Tatsuoka, 1971, pp. 133-134).

Furthermore, NR = ZZ', where Z is the standard score matrix, and since we are concerned only with the ranks of the matrices, the non-zero factor N is irrelevant and the rank of R equals the rank of Z (for example, see Harman, 1960, pp. 62-63).

---

[1]The author wishes to gratefully acknowledge the contribution of Mr. Nick L. Smith in the generation of this paper.

Finally, since the rank of a matrix is never greater than its smaller dimension (for example, see Horst, 1963, p. 334), the rank of Z is less than or equal to the smaller of N or V. Therefore, the maximum number of linearly independent factors that can be extracted from the correlation matrix R is less than or equal to the smaller dimension, N or V, of the original standard score matrix.

Humphreys (1964) in discussing Kaiser's rule of thumb for extracting only as many common factors equal to the number of roots greater than one of the complete correlation matrix (i.e., with ones in the main diagonal), suggests that with a small N, even this criterion might result in retaining a factor which is dependent only on chance. By using a small N, one may be capitalizing on sampling error in interpreting factors.

Aleamoni (1964) factor analyzed 66 observations on 62 variables using the Principal Axes procedure with Varimax rotation. Then, using a table of random numbers, he selected three subsamples of N = 51, N = 33, and N = 17, respectively, and attempted to factor analyze them. The subsample of 17 observations could not be factored, however, since the communalities were greater than one and were not acceptable. He attributes this to either the small N of 17 or else computer error.

The factor analysis of the subsamples of sizes 51 and 33 did produce interesting results, though. The subsample of N = 51 gave several factors quite similar to those of the total sample of N = 66. The subsample of N = 33, however, gave only two factors that were similar to those of the original sample. Aleamoni concluded that as N becomes less than or equal to V, the resultant intercorrelation matricies become less similar than those where N is larger than V.

Humphreys, et al., (1969) state that, as yet, no minimum N can be specified, but that N should be as large as feasible so that factors are based on stable differences among correlations as well as on correlations that are significantly greater than zero. They recommend including the smallest number of variables which will still serve the purposes of the investigation, and limiting the number of factors extracted to one-quarter of the number of variables. They further suggest that a trade-off between the number of observations, number of variables, and number of factors is a reasonable procedure. If, for example, only a limited number of observations is possible, then the number of variables studied and factors extracted should also be restricted.

Even when N is relatively large, however, extracted factors can still be due to chance. Using N's of 48, 96, and 384, and V's of 12, 24, and 48, Humphreys, et al., (1969) have been able to construct apparently well-defined factors from intercorrelations of random normal deviates. They state:

> *Empirically there are a great many factor-analytic investigations reported in which many of the variables have distributions of correlations that do not differ markedly from random distributions.* (p. 268)

This does not mean, of course, that all such factors are necessarily random, but only that better data are needed before one can confidently conclude that they are nonrandom. Of course, large sample sizes reduce the probability that factors are attributable entirely to chance.

Solomon (1966) followed this general approach of getting additional data to confirm the existance of factors in his study of teacher behavior dimensions. Solomon (Solomon, Bezdek, and Rosenberg, 1964; and Solomon, Rosenberg, and Bezdek, 1964) first of all factor analyzed 24 observations on 169 variables and extracted eight factors. He stated that he realized he had violated the

N greater than V requirement, but that this study was exploratory and that the number of variables was kept large "...so that the possibility of obtaining new and/or more subtle dimensions than have emerged previously would be maximized" (Solomon, Bezdek, and Rosenberg, 1964, p. 32). As has already been pointed out, however, with N less than V, it is the size of N that determines the number of possible dimensions and not V.

Solomon (1966), however, does go on to attempt to show that his factors are nonrandom. He factor analyzed 69 variables (items which were selected according to their loadings on the questionable factor analysis of the previous study) with a sample of 229 observations. Solomon reports that seven of the eight previous factors appeared again, in addition to four new ones.

This use of replication to show that results are nonrandom is, of course, a common and accepted practice. Cohen and Guthrie (1966), for example, in studying motivation patterns of college attendance, factor analyzed 105 variables on two samples of 105 and 95 observations, respectively. Although both sample sizes were probably much too small, they attempted to use the results of the smaller sample to confirm the results of the larger. They report that only six of the ten factors described in the first analysis were confirmed by the second analysis.

Although it appears on the surface that replications may help somewhat in confirming that factors defined from analyses with small N are nonrandom, Humphreys, et al., (1969) strongly caution:

> *Replicability, which is the mainstay of the scientific method, is hopeless in factor analysis studies unless hedged about with more controls than is commonly the case. It is clear that with appropriate values of N, n [number of variables], and m [number of factors extracted] the Procrustes method, either oblique or orthogonal, could replicate random factors endlessly.* (p. 269)

Thus, investigators seem to know that sample sizes should be larger than the number of variables before legitimately doing a factor analysis, but it appears from these studies that they frequently do not understand why this is the case and so tend to violate the constraint. Nor does replication seem to be an entirely satisfactory way to compensate for a small sample size.

The only recourse seems to be for investigators more strictly to adhere to the restriction of not using factor analysis unless the sample size is considerably larger than the number of variables. Even though previous investigators have stated that no minimum N can be specified, if one is interested in maximizing the number of possible dimensions underlying V, then N must be at least greater than V.

References

Aleamoni, L. M. Analysis of personality measures and measures used to predict success in science. Unpublished master's thesis, University of Utah, 1964.

Cohen, A. G., & Guthrie, G. M. Patterns of motivation for college attendance. *Educational and Psychological Measurement*, 1966, *26*(1), 89-98.

Guilford, J. P. *Psychometric methods*. New York: McGraw-Hill, 1947.

Horst, P. *Matrix algebra for the social scientist*. New York: Holt, Rinehart, and Winston, Inc., 1963.

Harman, H. H. *Modern factor analysis*. Chicago: University of Chicago Press, 1960.

Humphreys, L. G. Number of cases and number of factors: An example where N is very large. *Educational and Psychological Measurement*, 1964, *24*(3), 457-466.

Humphreys, L. G., Ilgen, D., McGrath, D., & Montanelli, R. Capitalization on chance in rotation of factors. *Educational and Psychological Measurement*, 1969, *29*(2), 259-271.

Murdoch, D. C. *Linear algebra for undergraduates*. New York: John Wiley & Sons, 1957.

Solomon, D. Teacher behavior dimensions, course characteristics, and student evaluations of teachers. *American Educational Research Journal*. 1966, *3*(1), 35-47.

Solomon, D., Bezdek, W., & Rosenberg, L. Dimensions of teacher behavior. *Journal of Experimental Education*, 1964, *33*(1), 23-40.

Solomon, D., Rosenberg, L. & Bezdek, W. Teacher behavior and student learning. *Journal of Educational Psychology*, 1964, *55*(1), 23-30.

Tatsuoka, M. M. *Multivariate analysis: Techniques for educational and psychological research*. New York: John Wiley & Sons, Inc., 1971.

Thurstone, L. L. *Multiple factor analysis*. Chicago: University of Chicago Press, 1947.